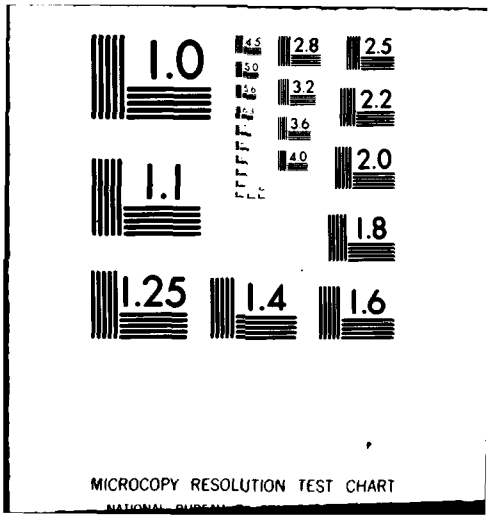
AD-A092 608 STANFORD UNIV CA DEPT OF STATISTICS F/G 12/1
AN APPROXIMATION TO THE DISTRIBUTION OF THE SAMPLE VARIANCE.(U)
OCT 80 H SOLOMON, M A STEPHENS N00014-76-C-0475
UNCLASSIFIED TR-291 NL

END
DATE FILMED
8-81
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART

(6)

AN APPROXIMATION TO THE DISTRIBUTION OF THE SAMPLE VARIANCE.

By

(10) Herbert Solomon and Michael A. Stephens

(11) 15 oct 80 (12) 45

(14) TR-

(9) TECHNICAL REPORT NO. 291

Accession For	
NTIS GRA&I	☒
DDC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist.	Avail and/or special
A	

October 15, 1980

(15) Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATSITICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

AN APPROXIMATION TO THE DISTRIBUTION OF THE SAMPLE VARIANCE

By

Herbert Solomon and Michael A. Stephens

1. INTRODUCTION.

In a recent paper, Tan and Wong (1977) investigated approximations to the distribution of the sample variance $s^2 = \Sigma(x_i-\bar{x})^2/(n-1)$ when the observations are drawn from parent populations which are not normal. In particular, they compared an approximation by Box (1953), using two moments of s^2, with a generalization of Box's approximation introduced by themselves, and with an approximation by Roy and Tiku (1962). The last two approximations use k moments of s^2, with $k \geq 4$, to approximate the density by a series expansion in Laguerre polynomials. The first k cumulants of s^2 and hence the first k moments, for any parent population with cumulants up to order $2k$, can be found by Fisher's k-statistics. Fisher (1928) gives the formulas for the first six cumulants of s^2 in terms of parent cumulants; the first four moments, both about the origin and about the mean, were given by Church (1925). For order higher than $k = 6$, in general, the cumulants or moments would be very complicated to calculate, and present obvious possibilities of error.

For comparing the approximations, Tan and Wong used as parent population a mixture of two normal distributions with the same variance, so that the mixing proportion p and the difference Δ between the means were two parameters which could be varied. By changing these and the sample size n very different distributions of s^2 can be produced;

for example, as Δ becomes larger the distribution of s^2 may become bimodal. Fortunately, for this parent population, Tan and Wong could compute the k-th moment of s^2 relatively easily for quite large k, and they also derived the exact density of s^2. Thus for this distribution exact probabilities could be compared with the approximations, with $k = 4$, 6 and 10 moments used for the Tan-Wong and Roy-Tiku approximations. Naturally the approximations using more moments were much better, with, on the whole, the Roy-Tiku approximation better than the Tan-Wong approximation when the same number of moments is used. However, the Roy-Tiku approximation can give negative values for the density, and Tan and Wong discuss this difficulty in some detail.

In this paper we give an approximation which has yielded good results. It uses only three moments of s^2, is relatively easy to compute, and never gives negative densities.

2. APPROXIMATION BY A GENERALIZED CHI-SQUARE DENSITY.

The approximation now suggested can be regarded as a generalization of Box's method. Box approximated the density of s^2 by $a\chi_b^2$, where a and b are constants determined by equating the first two moments of s^2 to those of $a\chi_b^2$. We now propose that the density of s^2 should be approximated by that of Y, where $Y = (cW)^k$ and where W has a χ^2 distribution with r degrees of freedom. This is written symbolically as $s^2 = (c\chi_r^2)^k$. The parameters c, r, and k are found by equating the first three moments of s^2, about the origin, to those of Y; the latter are given by $\mu(Y) = (2c)^k \Gamma(k+t)/A$, $\mu_2'(Y) = (2c)^{2k} \Gamma(2k+t)/A$, $\mu_3'(Y) = (2c)^{3k} \Gamma(3k+t)/A$, where $t = r/2$ and $A = \Gamma(t)$. Only the first three moments of s^2 are used to fit the approximate density; the fact that s^2 has a density with zero as its lower endpoint is being indirectly used also. The approximation Y has been employed with success by the authors for other random variables known to be positive (see, e.g. Solomon and Stephens, 1977) and the present application for the sample variance provides a natural situation to examine its usefulness once again.

3. ACCURACY OF THE APPROXIMATION.

Tan and Wong (1977) consider the distribution of s^2 for samples of size n from the mixed normal parent population with density

$$f(x) = p\ \phi_1(x;\mu_1,\sigma^2) + q\ \phi_2(x;\mu_2,\sigma^2) \qquad (1)$$

where $0 \leq p \leq 1$ and $q = 1-p$. In their Table 2, they give a comparison of exact values of the cumulative distribution of $Q = (n-1)s^2$ (although the table heading refers to s^2) with the Box approximation, their own approximation and the Roy-Tiku approximation: for the latter two approximations they use 4, 6, and 10 moments for s^2. In Table 1 we give exact values E, approximate values S given by the $(c\chi^2_r)^k$ approximation, and values $W = 10^4(S-E)$, of the cumulative distribution of Q, for a set of values of the parameters, and also of values of x, used by Tan and Wong. In the distribution (1) we have used $\sigma^2 = 4$, as did Tan and Wong.

In Table 2, values of W are recorded, together with values given by the Tan-Wong approximation (W_{TW}) and those given by the Roy-Tiku approximation (W_{RT}) when only 4 moments are used; these have been taken from Table 2 of Tan and Wong (1977). It can be seen that the new approximation S fares well on the whole. S is clearly unimodal, so cannot follow the density of s^2 when this tends towards or becomes bimodal - this is the case for low p and large Δ for example. S is then poor in the lower tail, but nevertheless gives good results in the upper tail which contributes most to the three moments used in the fit. Pearson curves were also fitted to the s^2 distribution, using either four moments or, when

possible, three moments and the lower endpoint (for this technique see Solomon and Stephens, 1978). These curves are also unimodal and give positive densities; however, sometimes the Pearson curve fit gives no lower endpoint, and overall it provides little or no improvement over the S approximation. The latter has the advantage of using only one χ^2_r distribution, and values of the cumulative χ^2_r are nowadays readily available from computer routines. Thus probabilities and percentage points for s^2 can be approximated easily once the fit has been made. A FORTRAN program is available for this purpose from the authors. We have noticed also that the fit will usually be very good if r is slightly altered, with the necessary adjustment made to c and k to match the first two moments; thus the cumulative distributions tabulated for r at intervals of 0.2 in Pearson and Hartley (1972) might also prove useful if r were adjusted to be one of the entries.

A further advantage of the proposed approximation is that it will be very easy to simulate the distribution of s^2 from any parent with at least six low-order cumulants.

Partially supported by the National Science and Engineering Research Council of Canada.

TABLE 1

$Q = (n-1)s^2$ is based on a sample of size n from density (1), with $\sigma^2 = 4$, and $\Delta = \mu_1 - \mu_2$. The table gives exact (E) values of $P(Q < X)$, values (S) obtained by the generalized chi-square approximation, and the difference $W = 10^4(S-E)$.

n = 3	r = 2.050	X: 0.3	1.0	3.0	4.0	8.0
p = .1	k = 1.018	E: .1293	.3963	.7481	.8404	.9740
Δ = 2	c = 1.041	S: .1290	.3694	.7481	.8404	.9740
		W: 3	1	0	0	0
n = 3	r = 1.149	X: 0.3	2.0	5.0	7.0	13.0
p = .1	k = 0.875	E: .1040	.4868	.7597	.8424	.9551
Δ = 6	c = 4.123	S: .1498	.4789	.7457	.8393	.9590
		W: 458	-79	-140	-31	39
n = 11	r = 10.309	X: 6.0	8.0	10.0	16.0	22.0
p = .1	k = 1.023	E: .1469	.3092	.4859	.8547	.9715
Δ = 2	c = 0.999	S: .1470	.3092	.4859	.8546	.9715
		W: 1	0	0	-1	0
n = 11	r = 3.930	X: 8.0	11.0	17.0	26.0	38.0
p = .1	k = 0.754	E: .1384	.2701	.5245	.8018	.9594
Δ = 6	c = 12.548	S: .1379	.2579	.5189	.8066	.9598
		W: -5	-122	-56	48	4
n = 11	r = 0.833	X: 5.0	11.0	35.0	53.0	77.0
p = .1	k = 0.370	E: .0342	.2035	.5759	.8259	.9697
Δ = 10	c = 30125	S: .0705	.1709	.5882	.8239	.9688
		W: 363	-326	123	-20	-9
n = 3	r = 1.896	X: 0.1	1.0	3.0	4.0	8.0
p = .4	k = 0.962	E: .0390	.3291	.7000	.8001	.9613
Δ = 2	c = 1.379	S: .0398	.3292	.6998	.8000	.9614
		W: 8	1	-2	-1	1
n = 11	r = 9.392	X: 7.0	11.0	15.0	19.0	23.0
p = .4	k = 0.954	E: .1518	.4522	.5994	.8818	.9559
Δ = 2	c = 1.498	S: .1518	.4521	.5994	.8818	.9559
		W: 0	-1	0	0	0
n = 11	r = 6.203	X: 27.0	33.0	39.0	45.0	50.0
p = .4	k = 0.609	E: .3638	.5818	.7631	.8828	.9406
Δ = 6	c = 49.80	S: .3667	.5820	.7611	.8812	.9400
		W: 29	2	-20	-16	-6
n = 11	r = 2.829	X: 47.0	71.0	83.0	89.0	107.0
p = .4	k = 0.309	E: .1127	.5258	.7579	.8457	.9797
Δ = 10	c = 432081	S: .1236	.5207	.7433	.8311	.9710
		W: 109	-51	-146	-146	-87

TABLE 2

Values of W (see Table 1) given by the new approximation, compared with values W_{TN} and W_{RT} given by the Tan-Wong and the Roy-Tiku approximations using only four moments.

	X:	0.3	1.0	3.0	4.0	8.0
n = 3	W:	3	1	0	0	0
p = .1	W_{TW}:	9	0	-4	0	1
Δ = 2	W_{RT}:	0	0	0	0	0

	X:	0.3	2.0	5.0	7.0	13.0
n = 3	W:	458	-79	-140	-31	39
p = .1	W_{TW}:	311	-82	-89	1	30
Δ = 6	W_{RT}:	-11	-7	14	7	-5

	X:	6.0	8.0	10.0	16.0	22.0
n = 11	W:	1	0	0	-1	0
p = .1	W_{TW}:	-1	-2	-1	1	0
Δ = 2	W_{RT}:	0	0	0	0	0

	X:	8.0	11.0	17.0	26.0	38.0
n = 11	W:	-5	-122	-56	48	4
p = .1	W_{TW}:	-25	-90	-3	36	-11
Δ = 6	W_{RT}:	-40	55	63	-65	15

	X:	5.0	11.0	35.0	53.0	77.0
n = 11	W:	363	-326	123	-20	-9
p = .1	W_{TW}:	65	-418	249	-110	-75
Δ = 10	W_{RT}:	-246	-609	-331	-217	129

	X:	0.1	1.0	3.0	4.0	8.0
n = 3	W:	8	1	-2	-1	1
p = .4	W_{TW}:	-14	-2	9	3	-3
Δ = 2	W_{RT}:	0	0	1	0	0

TABLE 2 (Continued)

	X:	7.0	11.0	15.0	19.0	23.0
n = 11	W:	0	-1	0	0	0
p = .4	W_{TW}:	2	3	-2	-2	0
Δ = 2	W_{RT}:	0	0	0	0	0
	X:	27.0	33.0	39.0	45.0	50.0
n = 11	W:	29	2	-20	-16	6
p = .4	W_{TW}:	33	-40	-51	-18	7
Δ = 6	W_{RT}:	114	29	-61	-82	-56
	X:	47.0	71.0	83.0	89.0	107.0
n = 11	W:	109	-51	-146	-146	-87
p = .4	W_{TW}:	240	-216	-182	-119	-24
Δ = 10	W_{RT}:	262	191	-274	-416	-342

References

Church, A.E.R. (1925). On the Moments of the Distribution of Squared Standard-Deviations for Samples of N Drawn from an Indefinitely Large Population. Biometrika, 17, 79-83.

Fisher, Ronald A. (1928). Moments and Product Moments of Sampling Distributions. Proc. Lond. Math. Soc. Ser. 2, 30, 199-238.

Pearson, E.S. and Hartley, H.O. (1972). Biometrika Tables for Statisticians Vol. 2. Cambridge: Cambridge University Press.

Roy, J. and Tiku, M.L. (1962). A Laguerre Series Approximation to the Sampling Distribution of the Variance. Sankhya, 24, 181-184.

Solomon, H. and Stephens, M.A. (1978). Approximations to Density Functions using Pearson Curves. Jour. Amer. Statist. Assn., 73, 153-160.

Tan, W.Y. and Wong, S.P. (1977). On the Roy-Tiku Approximation to the Distribution of Sample Variances from Nonnormal Universes. Jour. Amer. Statis. Assn., 72, 875-880.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 291	2. GOVT ACCESSION NO. AD-A092	3. RECIPIENT'S CATALOG NUMBER 608
4. TITLE (and Subtitle) AN APPROXIMATION TO THE DISTRIBUTION OF THE SAMPLE VARIANCE		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Herbert Solomon and Michael A. Stephens		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042 267
11. CONTROLLING OFFICE NAME AND ADDRESS OFFICE Of Naval Research Statistics & Probability Program Code 436 Arlington, VA 22217		12. REPORT DATE October 15, 1980 13. NUMBER OF PAGES 9
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Generalized chi-square approximation; Sample variance.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A simple three-moment approximation is introduced for the distribution of the sample variance. Comparisons are given with other approximations discussed by Tan and Wong (1977).

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)